



**Daffodil**  
*International*  
**University**

## **Analysis and Prediction of Cholera Disease using Machine Learning Algorithms**

**Submitted by**

Roisujaman Shabab

ID: 161-35-1603

Department of Software Engineering

Daffodil International University

**Supervised by**

Ms. Farzana Sadia

Assistant Professor

Department of Software Engineering

Daffodil International University

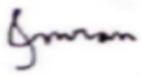
This Thesis report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering.

© All right Reserved by Daffodil International University

## APPROVAL

This Thesis titled on “Analysis and Prediction of Cholera Disease using Machine Learning Algorithms”, submitted by Roisujaman Shabab, ID: 161-35-1603 to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

## BOARD OF EXAMINERS



---

Dr. Imran Mahmud  
Associate Professor and Head  
Department of Software Engineering  
Daffodil International University

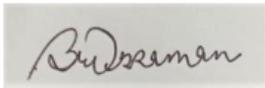
Chairman



---

Syeda Sumbul Hossain  
Senior Lecturer  
Department of Software Engineering  
Daffodil International University

Internal Examiner 1



---

Khalid Been Badruzzaman Biplob  
Senior Lecturer  
Department of Software Engineering  
Daffodil International University

Internal Examiner 2



---

Professor Dr. Mohammed Nasir Uddin  
Department of Computer Science and Engineering  
Jagannath University, Dhaka

External Examiner

## Thesis Declaration

I hereby declare that thesis title “Analysis and Prediction of Cholera Disease using Machine Learning Algorithms” has been completed by me under the supervision of Ms. Farzana Sadia, Assistant Professor, Department of Software Engineering, Daffodil International University for the purpose of achieving degree of Bachelor of Science from Daffodil International University. This is also declared by me that neither this thesis nor any part of this thesis has been used or submitted elsewhere for any kind of degree or awards.



---

**Roisujaman Shabab**

ID: 161-35-1603

Department of Software Engineering  
Daffodil International University



---

**Ms Farzana Sadia**

Assistant Professor

Department of Software Engineering  
Daffodil International University

**Supervisor**

## **ACKNOWLEDGEMENT**

I am grateful to my creator for allowing me to complete this research work and learn so much. I am thankful to my research supervisor, Ms Farzana Sadia, Assistant Professor, SWE, to provide careful guidance, starting from selecting the research scope to finalise the research work successfully. I would also like to thank Mr. Md. Anwar Hossen, Assistant Professor, SWE, for his valuable comments, which was always insightful. I am also thankful to all the lecturers, Department of Software Engineering, who sincerely guided me at my difficulty. I am grateful to my parents for their unconditional support and encouragement. I am thankful to my friends who supported me throughout this venture.

## Table of Contents

THESIS DECLARATION .....	i
ACKNOWLEDGEMENT .....	ii
Table of Contents .....	iii
LIST OF TABLES.....	vi
LIST OF FIGURE S .....	vii
LIST OF ABBREVIATIONS.....	viii
ABSTRACT.....	ix
CHAPTER 1 .....	1
INTRODUCTION .....	1
1.1 Background .....	1
1.2 Motivation of the Research .....	2
1.3 Problem Statement .....	2
1.4 Research Questions .....	2
1.5 Research Objective.....	3
1.6 Research Scope .....	3
1.7 Thesis Organization.....	3
CHAPTER 2 .....	4
LITERATURE REVIEW .....	4
2.1 Previous literature .....	4
2.2 Previous research on Analysis Effectiveness in Determining the Epidemic .....	4
2.3 Previous research on forecasting Cholera disease .....	5
2.4 Research Gap .....	8
2.5 Summary .....	11
CHAPTER 3 .....	8
RESEARCH METHODOLOGY .....	12
3.1 Research Dataset .....	12

3.2 Data Preprocessing.....	12
3.3 Algorithm Selection .....	14
CHAPTER 4 .....	11
RESULTS AND DISCUSSIONS .....	16
4.1 Data Analysis & Experimental Result .....	16
4.1.1 Experimental Result & Evaluation Matrix .....	16
4.1.2 Exploratory Data Analysis.....	18
CHAPTER 5 .....	23
CONCLUSIONS AND RECOMMENDATIONS .....	23
5.1 Findings and Contributions.....	23
5.2 Limitations .....	24
5.3 Recommendations for Future Works.....	24
REFERENCES .....	24

## LIST OF TABLES

Table 1: Literature Review .....	8
Table 2: Accuray Report of Gradient Boosting Algorithm.....	18
Table 3: Descriptive statistics of the dataset .....	

## LIST OF FIGURE S

Figure 1: Architecture Diagram of Proposed Research .....	12
Figure 2: Visualisation of data before cleaning .....	13
Figure 3: Visualisation of data after cleaning .....	14
Figure 4: Exploratory data analysis for the top 10 countries .....	21
Figure 5: Finding the correlation between the independent variable.....	22
Figure 6: Investigating the average number of deaths from Cholera.....	23
Figure 7: Investigating the Cholera status of Bangladesh .....	24

## ABSTRACT

In this research, we have investigated Cholera disease and its fatality rate from previous years in terms of various countries. This disease is not new, yet researchers are currently utilising several approaches to detect the difficulties and extract hidden information from previous records. This study proposed an alternative solution in terms of Cholera disease. Several traditional machine learning algorithms are experimented with to analyse and predict the disorder from the existing dataset. The proposed research methodology has consisted of four phases, for instance, Research Dataset, Data Preprocessing and Algorithm Selection. Our investigation shows that the Gradient Boosting algorithm performs well in this type of dataset with an accuracy of 93%. We have discovered the R2 score, Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE) are 0.932841%, 398.1827%, 158549.4724%, and 71.6621098%, respectively.

We have carried out an exploratory data analysis where Cholera case data analysis of Bangladesh has been done from 1996 to 2000. In addition to the disease prediction, data analysis of different countries has been carried out, and correlations have been made through which the interrelationships of each indicator can be found very quickly.

**Keywords:** Machine Learning (ML), RMSE, MSE, Gradient Boosting and Cholera Disease

# CHAPTER 1

## INTRODUCTION

### 1.1 Background

Cholera disease is not only a global issue but also a historical issues in the world. Cholera is an acute diarrhoeal infection caused by the bacterium *Vibrio Cholera* ingestion of infected food or water. Cholera remains a global public health issue and a symbol of inequity and a lack of social growth (WHO, 2021). This disease is significant and responsible for the various problem of our social life. According to the recent statistics, the scientist reported that every year 1.3 to 4.0 million Cholera case fatality rate of every year is due to less awareness and lack of self-consciousness.

In the context of Bangladesh, In 1991, a massive diarrhoea epidemic erupted in Bangladesh. To estimate the magnitude of Cholera during diarrhoea epidemics and to focus on Cholera-related public health problems in Bangladesh. Here, two types of people can be found in Bangladesh: the people who live in the urban area and another in the rural area. It is noticeable that those who lived in the urban area are aware than the people who stay in the village side. The risk of Cholera has increased with age, occupation and recent history of diarrhoea among family members. Conclusion: Cholera is found in large part of Bangladesh. Cholera-prone areas should be prioritised in disease control through the implementation of targeted interventions.

## **1.2 Motivation of the Research**

Cholera is a significant disease, and it is essential to do awareness and analysis among people. Medical data analysis is relatively difficult, so it is crucial to analyse it through computational techniques. Machine Learning (ML) can diagnose diseases more effectively at an earlier level, reducing the number of readmissions in hospitals and clinics. Technology has also advanced significantly in discovering and developing new medicines with tremendous potential for treating patients with complex conditions. So, it is a matter of great concern to bring an effective solution and analyse the difficulties. This proposed study's motivation is to utilise the ML approach predicting the Cholera case fatality rate from the existing dataset and Figure out meaningful insights.

## **1.3 Problem Statement**

We have found that the previous research has not been accomplished or adequately formulated. Most of the previous research has been done through statistical analysis, but very little work has been done on Cholera disease prediction using machine learning algorithms. We also noticed that no detailed pipeline on data processing had been provided in previous research, which has caught our attention. Many evaluation matrixes show how good the model is, which has not been shown in previous research.

## **1.4 Research Questions**

The research question was

- RQ1: Which algorithm performs best in predicting Cholera disease?
- RQ2: Is there only one or multiple variables that positively influence the disease prediction?

## **1.5 Research Objective**

The following objectives are addressed in this proposed study:

- To find out the best algorithm that performs well in terms of Cholera disease prediction on Cholera cases population demographic data..
- Analyzing Cholera Population on Demographic Data
- To extract meaningful insights from the previous Cholera cases records.
- To achieve promising accuracy by experimenting with various algorithms.
- To reduce the loss function as well as overfitting and assure the model's generalizability.

## **1.6 Research Scope**

We have carefully investigated the Cholera disease and predict using the previous dataset. Since we have researched the clinical aspect, it will be possible to identify the Cholera disease fatality rate through our model accurately as well as find out what kind of factor works with this disease. We have shown the correlation to the interrelation of each feature so that it will serve as a benchmark in the research community.

## **1.7 Thesis Organisation**

In the following, in second chapter we examined about different looks researches done on the same subject including the research gap. In chapter three, we examined our proposed research methodology. In chapter four, we discussed the analysis results. Finally, in chapter five, we discussed the findings, limitations and future work.

## CHAPTER 2

### LITERATURE REVIEW

#### 2.1 Previous literature

In this research (Emch et al., 2008), the environmental conditions have been linked to Cholera and thus may help predict the disease. Compared to satellite-derived and in-situ ecological time-series data in Bangladesh and Vietnam, temporal Cholera distributions are compared to satellite-derived and in-situ ecological time-series data to investigate the links between Cholera and the local climate. In Bangladesh, ordered probit models are used to look at associations, while in Vietnam, probit models are used to look at associations at two different locations. Increased levels of ocean chlorophyll are linked to a rise in the severity of Cholera in Bangladesh. Increases in sea surface temperature. In Hue, Vietnam, the weather has the most significant impact, while in Nha Trang, Vietnam, river height has a significant effect.

#### 2.2 Previous research on Analysis Effectiveness in Determining the Epidemic

Epidemic disease outbreaks have prompted today's population to become increasingly concerned with infectious disease control, prevention, and management strategies to reduce disease spread and infected areas. For the contagious disease countermeasure and prediction study, the backpropagation approach was used in this research (Ibrahim, Akhir, & Hassan, 2017). Machine learning can evaluate the predictive analysis based on the backpropagation approach, which encourages artificial intelligence in pattern recognition, statistics, and feature selection. The feature collection of disease transmission factors that are likely to have strong interconnections in causing infectious disease outbreaks is the classification technique.

The spread of Cholera is modelled using a spatially explicit scheme that considers the dynamics of prone, contaminated, and recovered individuals in various local communities linked by hydrologic and human mobility networks. The study (**Pasetto, Finger, Rinaldo, & Bertuzzo, 2017**) introduces two major Cholera modelling innovations: using a data assimilation technique, specifically an ensemble Kalman filter, to change state variables and parameters based on observations and the use of rainfall forecasts to force the model. A benchmark was developed by simulating the state of the system and the predictive capabilities of novel tools during the early stages of the 2010 Haitian Cholera outbreak using only knowledge available at the time. Their findings indicate that the assimilation protocol with sequential parameter updates outperforms Markov chain Monte Carlo calibration schemes.

**In (Singh, Singh, & Bhatia, 2018)**, the authors addressed that datasets available on the internet containing helpful information are known as sentiment analysis (SA). In the field of computer science, machine learning techniques play a significant role. The study of patterns and the creation of computational systems that can learn and make predictions is one of the applications of machine learning techniques. There have been studies in the recent past. Data sets were taken from microblogging websites such as Twitter, and essential healthcare information was produced through sentiment analysis of data sets. The importance of sentiment analysis in predictions is described; (ii) the relevance of sentiment analysis using Machine Learning (ML) techniques is studied; (iii) data classification using ML techniques is also described, and (iv) a survey on the use of microblogging sites to predict outbreaks and epidemics is conducted through some major research articles from 2010 to 2017.

A seasonal-auto-regressive integrated-moving-average (SARIMA) model was used for time-series analysis between 2000 and 2013 due to the auto-regressive nature of Cholera and its seasonal

behavioural trends (**Daisy et al., 2020**). Since rainfall ( $r = 0.43$ ) and maximum temperature ( $r = 0.56$ ) have the most significant impact on Cholera incidence, single-variable (SVMs) and multi-variable (MVMs) SARIMA models (MVMs) were established, compared, and tested to determine their relationship with Cholera incidence. Relative humidity ( $r = 0.28$ ), ENSO ( $r = 0.21$ ), and SOI ( $r = -0.23$ ) were all found to have a weak relationship. A 7% rise in Cholera incidence ( $p = 0.001$ ) was observed using SVM for a 1 °C increase in maximum temperature at a one-month lead time. MVM, on the other hand, outperformed SVM (AIC 14.15, BIC 14.36) in terms of efficiency.

### **2.3 Previous research on forecasting Cholera disease**

The authors (**Daisy et al., 2020**) present a novel exploration of the potential of a machine learning method to forecast environmental Cholera risk in coastal India, which has a population of more than 200 million people, using critical climate variables derived from atmospheric, terrestrial, and oceanic satellites. A Random Forest classifier model is created, educated, and tested on a Cholera outbreak dataset for districts along the coast of India from 2010 to 2018. With an Accuracy of 0.99, an F1 Score of 0.942, and a Sensitivity score of 0.895, the random forest classifier model correctly identifies 89.5 percent of outbreaks. Seasons and coastal locations revealed spatial-temporal trends in the model's results.

The authors (**Badkundri, Valbuena, Pinnamareddy, Cantrell, & Standeven, 2019**) proposed the Cholera Artificial Learning Model (CALM), a series of four extreme-Gradient boosting (XGBoost) machine learning models that forecast the number of new Cholera cases a Yemeni governorate will encounter over a period ranging from two weeks to two months. CALM uses rainfall data, past Cholera cases and deaths data, civil war casualties, and inter-governorate experiences from different time frames to create a novel machine learning approach. CALM may also teach complex nonlinear relations that appear in epidemiological phenomena thanks to

machine learning and comprehensive function engineering. In a real-world simulation, CALM can predict Cholera incidence 2 weeks to two months ahead of time with a margin of 5 Cholera cases per 10,000 people.

The study (**Leo, Luhanga, & Michael, 2019**) proposed Machine learning techniques to model Cholera epidemics with seasonal weather shifts, thus solving the data imbalance problem, are explored in this paper. The dataset's sampling equilibrium and dimensionality were restored using the Adaptive Synthetic Sampling Approach (ADASYN) and Principal Component Analysis (PCA). The performance of the seven models was also assessed using sensitivity, specificity, and balanced accuracy metrics. XGBoost classifier was chosen as the best model for the analysis based on the Wilcoxon sign-rank test results and model features. Overall, the findings helped us better understand the critical functions of machine learning techniques in healthcare data.

Early identification of Cholera cases is critical for limiting the severity and length of an outbreak. However, by the time Cholera cases are reported, an epidemic could already be well underway. This paper explains how to use data from various sources and machine learning techniques to forecast the likelihood of Cholera outbreaks in different areas over time. The authors examine regions with similar Cholera prevalence dynamics over time using data collected across 80 Ugandan regions. They then develop a probabilistic model to forecast possible Cholera cases and explain how it works (**Mubangizi, Mwebaze, & Quinn, 2009**).

The authors (**Chau et al., 2017**) of this paper suggested a new approach to improve cholera outbreak prediction efficiency in Hanoi, Vietnam. Solar terminology, training data resampling, and classification methods are all included in the new approach. The results of the experiments show that combining solar terms with ROSE resampling and the random forests method results in (AUC) with balanced sensitivity and specificity.

## **2.4 Research Gap**

Based on the review of the above research, it can be said that most of the research has been completed with analysis but using machine learning technology in the context of Disease Prediction and Bangladesh. By looking at the above literature review, it is obvious that majority of the research concentrated on statistical analysis or time series data handling for forecasting the outbreak, but to the best of our knowledge, there is not adequate study carried out in terms of cholera disease cases fatality rate prediction using machine learning approach. In this research, we have come up with an effective solution for combacting the issues. Therefore, by using our model it is possible to predict the fatality rate of cholera disease through the computational approach. On the most the cases, researcher often use their own strategy for solving this issues, but it is a matter of great concern that analyzing the disease in the context of Bangladesh is crucial that cannot be found beforehand in the previous study.

## **2.5 Summary**

A variety of machine learning algorithms have been used in our research, and Cholera Disease has been identified by choosing the best algorithm from there. We have extracted information from the previous data through data analysis technique in Bangladesh, which has not been extensively completed during the existing research.

## CHAPTER 3

### RESEARCH METHODOLOGY

Figure 1 illustrates the architecture diagram of the proposed research. By looking at Figure 1, it can be clearly observed that the architecture is consisted of several phases, for instance, Data collection, Data Preprocessing, Data Visualization, Extract Information, Algorithm Selection, Model Evaluating and Classifier. Hence, in this way, the disease has been classified and analyse the hidden information; however, the entire research is concentrated by following this procedure.

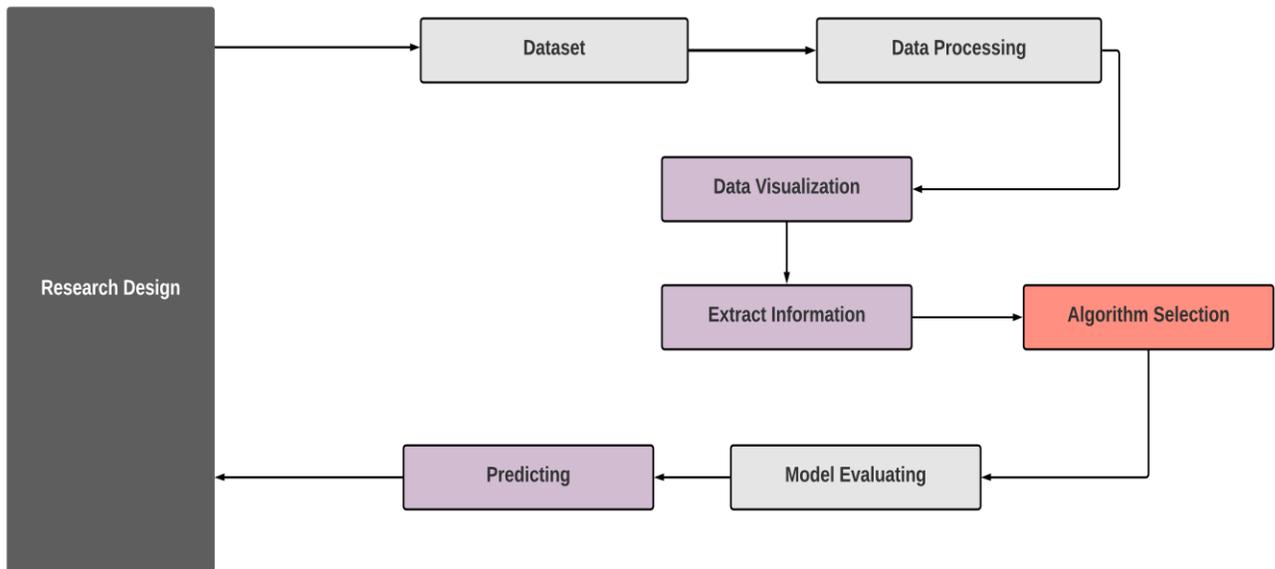


Figure 1: Architecture Diagram of Proposed Research

### 3.1 Research Dataset

We have used this dataset from the online repository (kp, 2020). This dataset Contains country-wise no. of cases, deaths and CFR (case fatality ratio) from the year 1949 till 2016. The number of reported deaths from Cholera can be found around 2492, and the Cholera case fatality rate is around 2492.

### 3.2 Data Preprocessing

The main goal of Data Cleaning is to find and delete errors and duplicate data so that a consistent dataset can be created. This increases the training data quality for analytics and allows for more precise decision-making. In this step, we have removed the null value from our dataset since it can be responsible for many issues; however, it has to be solved in order to feed the data into the machine learning algorithm. Figure 2 and Figure 3 shows the noise linked dataset illustration and cleaning dataset visualisation.

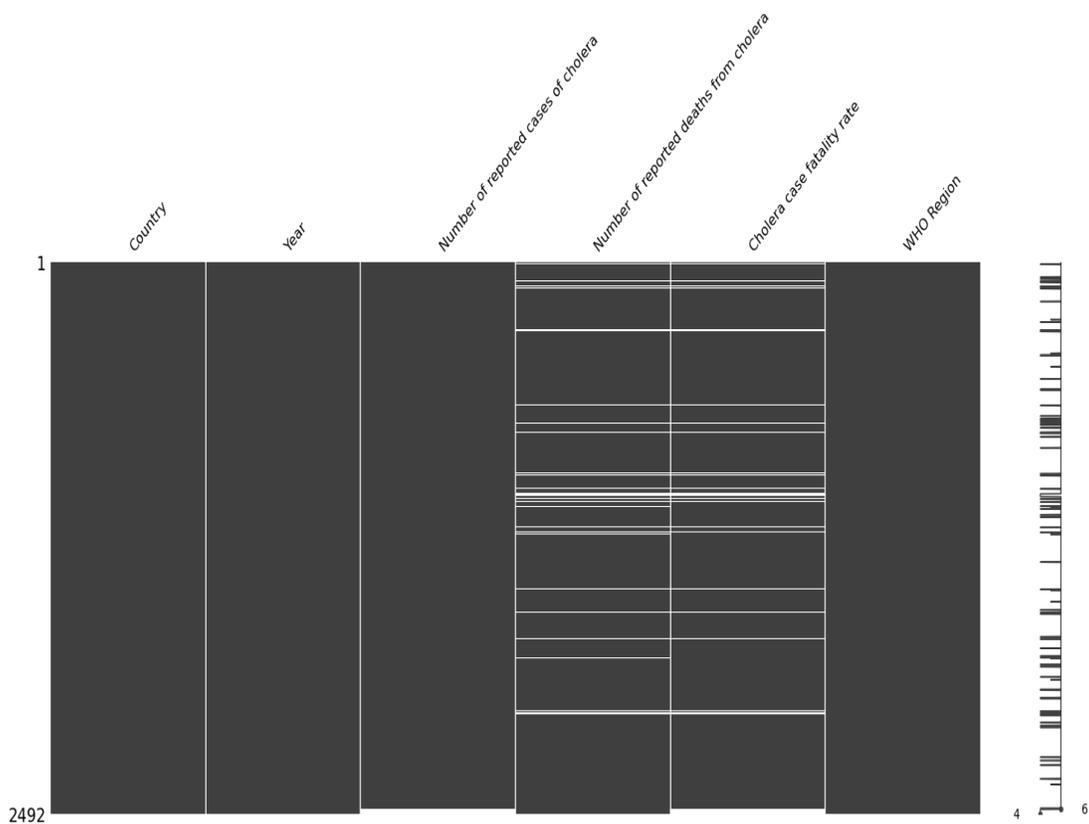


Figure 2: Visualisation of the noisy data before cleaning

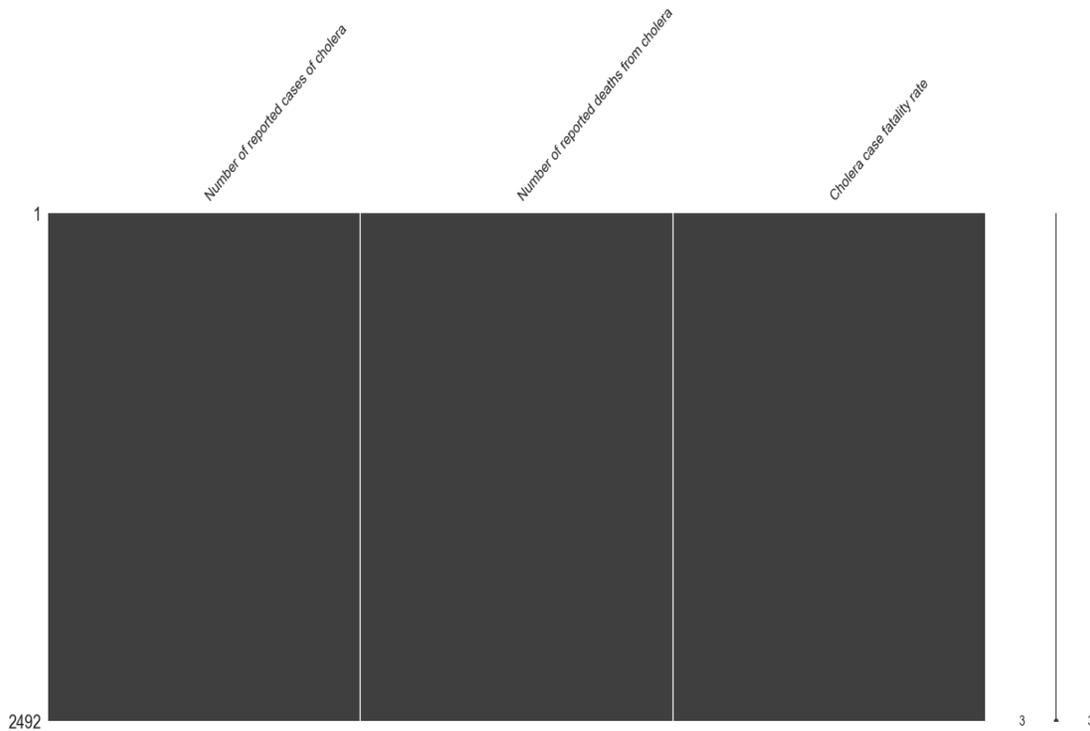


Figure 3: Visualisation of the data after cleaning

### 3.3 Algorithm Selection

We have used seven popular machine learning algorithms, for instance, Linear Regression, Logistic Regression, Support Vector Machine (SVM), Adaptive Boosting, Gradient Boosting, K nearest neighbors (KNN) and Decision Tree. In this section, the Gradient Boosting algorithm will have been interpreted because we have found satisfactory accuracy on this approach in terms of the Cholera disease prediction.

**Gradient Boosting:** The Gradient boosting algorithm also follows the sequential ensemble learning method. Through loss optimisation in this way, WikiLearners gradually became better than its previous weak learners. For example, the 2nd weak learner is better than the 1st. The 3rd weak learner is better than the 2nd, so as the weak learner periodicity increases, the amount of

error in the model decreases, and the model becomes a stronger learner. The Gradient boosting algorithm works relatively well for regression type problems (**González-Recio, Jiménez-Montero, & Alenda, 2013**).

The difference between Gradient boosting and Ada boosting is that in Adaptive boosting, error is gradually reduced by updating the weight of the wrong predictive air sample. In Gradient, boosting the loss function is optimised, and each loss is optimised. The amount of error also decreases. To optimise this loss function, each weak learner changes its alternative weak learner model so that the next weak learner is better than the previous one. On the other hand, Gradient boosting consists of 3 components, weak learner, loss function optimisation and additive model. The following equation (1), (2), (3), (4), (5) and (6) are shown mathematically the working procedure of the Gradient boosting algorithm.

*Step1:* Initialise the function estimate with a constant value  $\hat{f}(x) = \hat{f}_0, \hat{f}_0 =$

$$\gamma, \gamma \in \mathbb{R}, \hat{f}_0 = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma) \quad (1)$$

*Step2:* For each iteration  $t = 1, \dots, T$  :

i. Calculate pseudo-residuals  $r_t, r_{it} = - \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=\hat{f}(x)}$ , for  $i =$

$$1, \dots, n \quad (2)$$

ii. Add a new function  $g_t(x)$  (it can be any model, but here we are using decision trees) as regression on pseudo-residuals  $\{(x_i, r_{it})\}_{i=1, \dots, n}$  (3)

iii. Find optimal coefficient  $\rho_t$  at  $g_t(x)$  regarding initial loss function  $\rho_t =$

$$\arg \min_{\rho} \sum_{i=1}^n L(y_i, \hat{f}(x_i) + \rho \cdot g_t(x_i, \theta)) \quad (4)$$

iv. Update current approximation  $\hat{f}(x)$  where  $\hat{f}_t(x) = \rho_t \cdot g_t(x)$

$$\hat{f}(x) \leftarrow \hat{f}(x) + \hat{f}_t(x) = \sum_{i=0}^t \hat{f}_i(x) \quad (5)$$

Step 3: The final GBM model will be the sum of the initial constant and all the subsequent function updates  $\hat{f}(x) = \sum_{i=0}^T \hat{f}_i(x)$  (6)

## CHAPTER 4

### RESULTS AND DISCUSSIONS

#### 4.1 Data Analysis & Experimental Result

The result analysis section is subdivided into three segments: Experimental Result & Model Evaluation, Exploratory Data Analysis and Comparative Analysis.

##### 4.1.1 Experimental Result

To explain how well the model is doing in its predictions, evaluating the model accuracy is an integral part of developing machine learning models. The evaluation metrics differ depending on the type of problem. The linear model (regression) is a typical example of this type of problem, and the key feature of the regression problem is that the dataset's goals only include real numbers. The errors indicate how much the model makes errors in its predictions. The basic idea behind accuracy assessment is to equate the original target with the forecast one based on criteria. We experimented with the linear regression model in this dataset, but we did not find satisfactory accuracy; in the same dataset, the Gradient boosting algorithm has been applied, and good accuracy has been achieved. Table 1 shows the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R<sup>2</sup> and Accuracy. The MSE, MAE, RMSE, and R-Squared metrics are commonly used in regression analysis to assess prediction error rates and model efficiency. Figure 4 shows the accuracy graphs in terms of the algorithms that have been experimented while conducting this research towards Cholera diseases prediction.

Table 1: Accuracy report of the Gradient boosting algorithm

MAE %	MSE%	RMSE%	Accuracy%	R2%
71.6621098	158549.4724	398.1827	93.284	0.932841

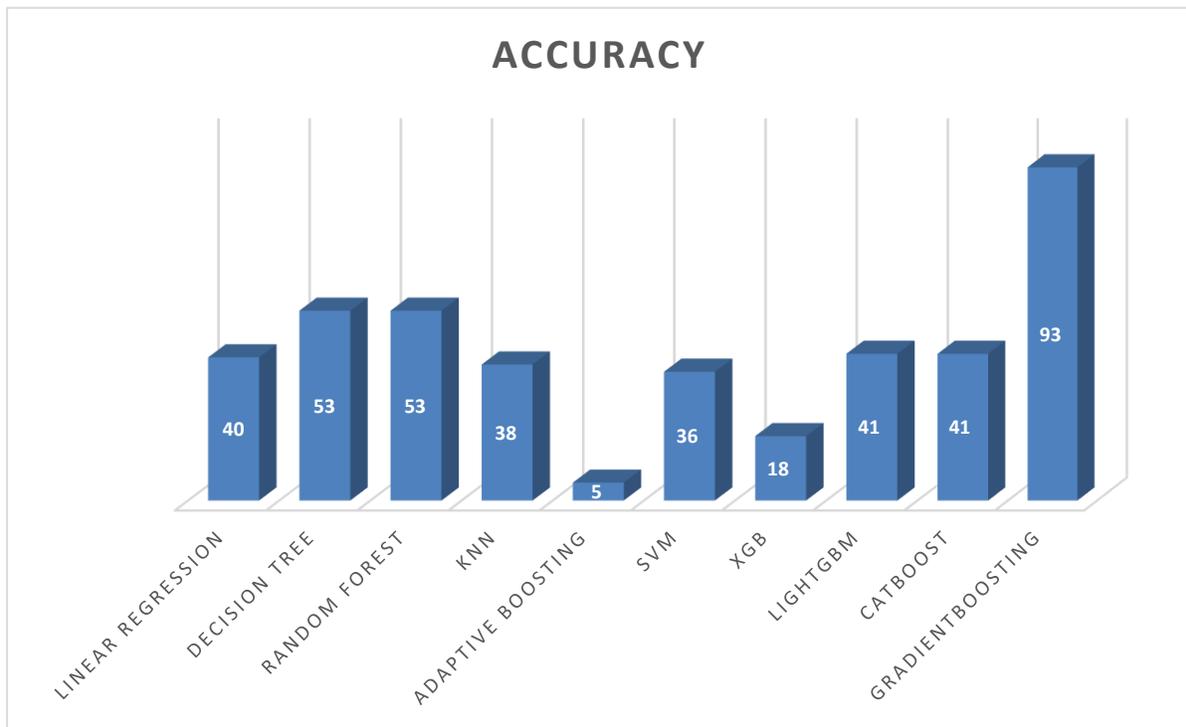


Figure 4: Accuracy Graphs of the algorithms that have been applied

#### 4.1.2 Evaluation Matrix

This section has been divided into several phases: the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared.

##### A. MAE (Mean absolute error):

It represents the difference between the original and expected values, calculated by averaging the fundamental difference over the entire data set.

### **B. MSE (Mean Squared Error):**

It represents the difference between the original and expected values calculated by dividing the average difference over the data set by square.

### **C. RMSE (Root Mean Squared Error):**

The square root of MSE divided by the error rate. The coefficient of determination (R-squared) represents how well the values match together.

### **D. R-squared (Coefficient of determination):**

It represents the coefficient indicating how well the values match together compared to the original values. The value ranges from 0 to 1 and is expressed as a percentage. The better the model, the higher the value. The equation (7), (8), (9) and (10) are used to find out the prediction error and model efficiency.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}| \quad (7)$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2 \quad (8)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2} \quad (9)$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} \quad (10)$$

Where

$\hat{y}$  – predicted value of  $y$   
 $\bar{y}$  – mean value of  $y$

### 4.1.3 Exploratory Data Analysis

This section analyses the Cholera case on top of different countries. Without analysing a stool sample, it is almost impossible to differentiate a single patient with Cholera from a patient infected with another pathogen that causes acute watery diarrhoea. Because of the disease's rapid dissemination, a study of clinical characteristics of several patients who are part of a suspected outbreak of acute watery diarrhoea may help to detect Cholera. While the treatment of patients with acute watery diarrhoea is the same regardless of the condition, identifying Cholera is critical due to the risk of a widespread outbreak. Bangladesh only has data from the year 2010 onwards. Despite a slight rise in the number of Cholera cases, we can see that Bangladesh has made progress in its Cholera war. This data set has some problems with data quality, such as missing values, data form mismatches, and invalid number input. Which ones were already corrected during the analysis? We can see that the number of outliers is minimal. We also discovered that not all countries have data for all years. The most significant number of confirmed deaths in 2016 was in the Democratic Republic of the Congo, Somalia, Haiti, the United Republic of Tanzania, Yemen, South Sudan, Kenya, Malawi, Nigeria, and the Dominican Republic. The countries with the most Cholera cases in 2016 were Haiti, the Democratic Republic of the Congo, Yemen, Somalia, the United Republic of Tanzania, Kenya, South Sudan, Malawi, the Dominican Republic, and Mozambique. The countries with the highest fatality rates in 2016 were Niger, Congo, Zimbabwe, Nigeria, Angola, Somalia, the Democratic Republic of the Congo, Malawi, Dominican Republic, and Uganda.

The total number of outbreaks, average death, and fatality rate have all decreased over time, but Cholera disease plagues a few countries. To eliminate Cholera, we need to focus more on these countries. The details sequence and consequence are shown separately in Figure 5,6,7, and 8.

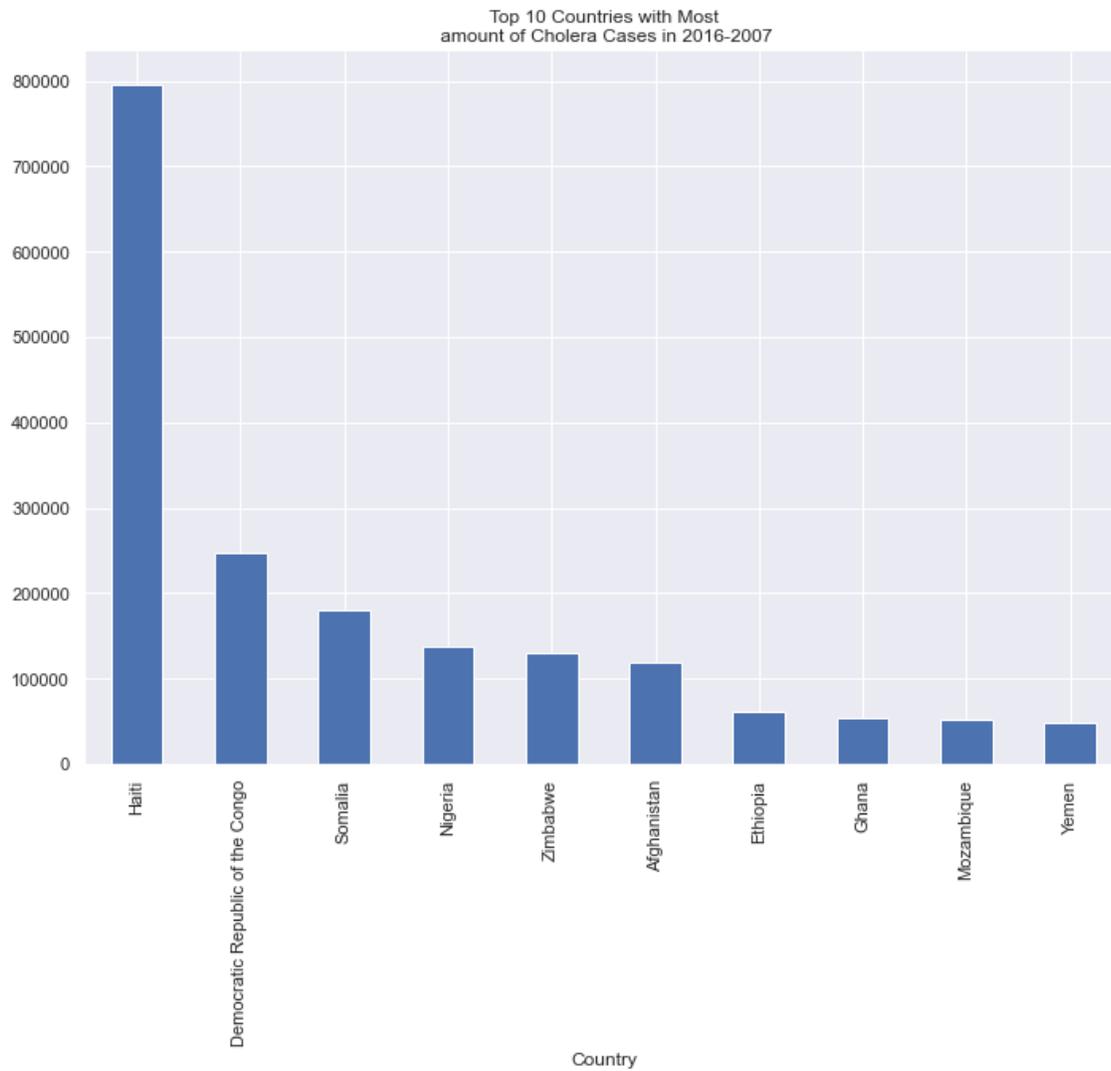


Figure 5: Exploratory data analysis for the top 10 countries with the most amount of Cholera disease.

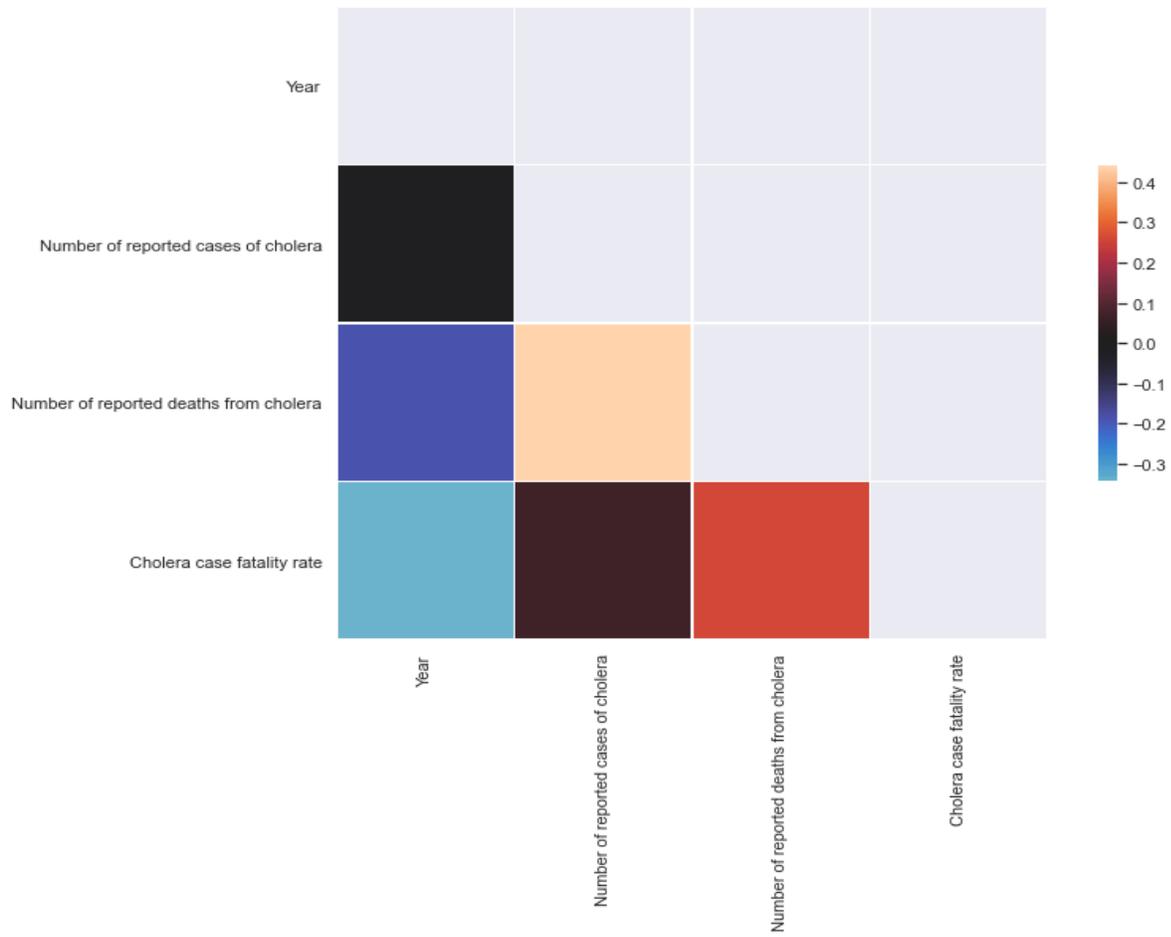


Figure 6: Finding the correlation between the independent variables in terms of Cholera disease.

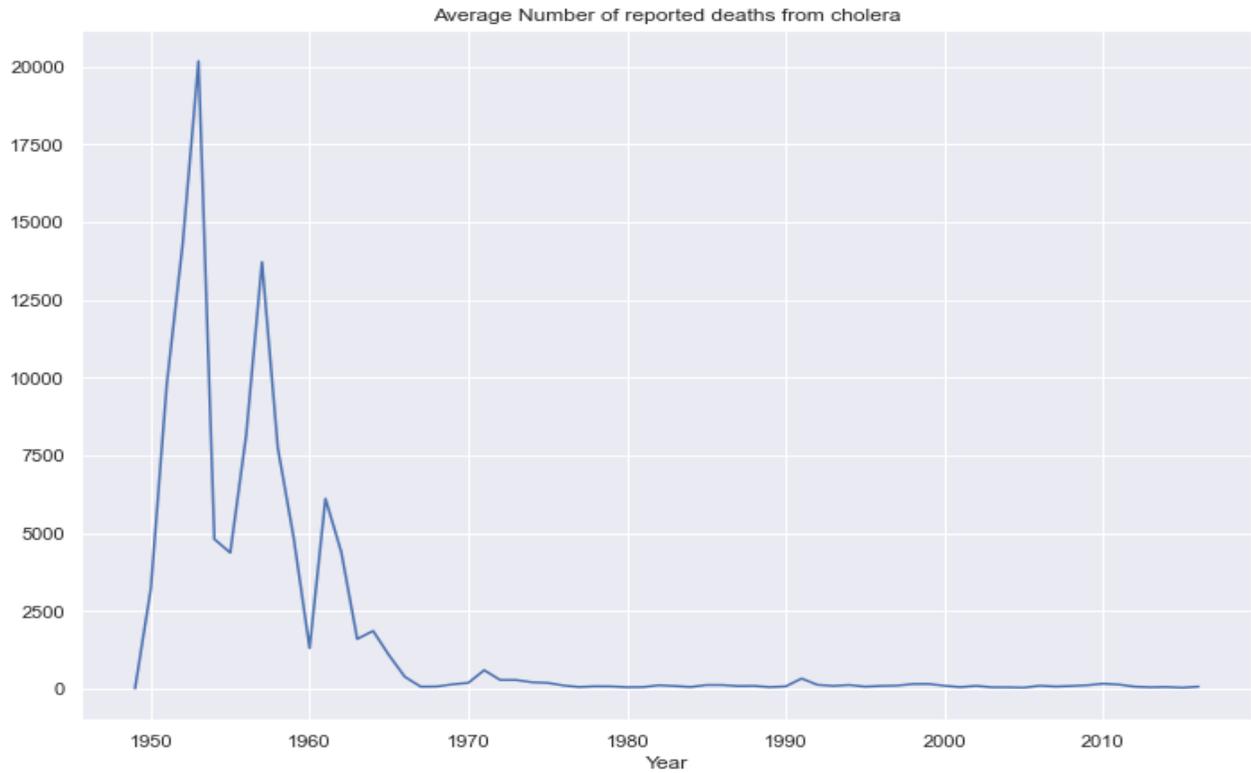


Figure 7: Investigating the average number of deaths from Cholera from 1950 to 2010

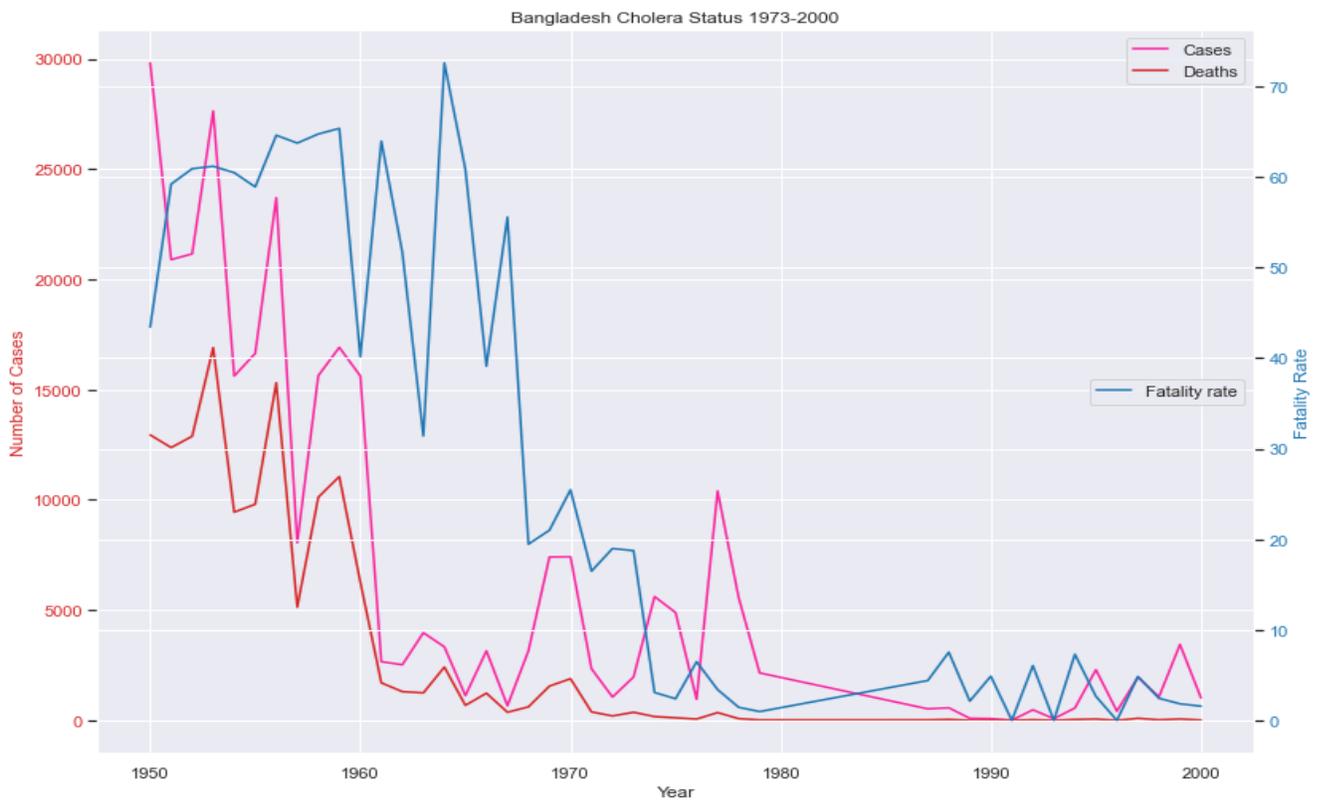


Figure 8: Investigating the Cholera status of Bangladesh from 1973 to 2000

The total number of outbreaks, average death, and fatality rate have all decreased over time, but Cholera disease plagues a few countries. To eliminate Cholera, we need to focus more on these countries. The details sequence and consequence are shown separately in Figure 2,3,4, and 5.

Table 2: Statistical analysis & measurements of the dataset

	<b>Year</b>	<b>Number of reported cases of Cholera</b>	<b>Number of reported deaths from Cholera</b>	<b>Cholera cases fatality rate</b>
<b>Count</b>	2492.000000	2492.000000	2492.000000	2492.000000
<b>Mean</b>	1992.343098	3684.060193	360.033708	5.459960
<b>Std</b>	14.834151	14840.198322	3484.892806	15.211705
<b>Min</b>	1949.000000	0.000000	0.000000	0.000000
<b>25%</b>	1981.000000	8.750000	0.000000	0.000000
<b>50%</b>	1994.000000	228.500000	5.000000	1.300000
<b>75%</b>	2004.000000	1847.750000	53.250000	4.912500
<b>max</b>	2016.000000	340311.000000	124227.000000	450.000000

Table 3: Correlation matrix of the Cholera disease attributes

	<b>Year</b>	<b>Number of reported cases of Cholera</b>	<b>Number of reported deaths from Cholera</b>	<b>Cholera case fatality rate</b>
<b>Year</b>	1.000000	-0.011789	-0.184387	-0.342870
<b>Number of reported cases of Cholera</b>	-0.011789	1.000000	0.442026	0.069941
<b>Number of reported deaths from Cholera</b>	-0.184387	0.442026	1.000000	0.258984
<b>Cholera case fatality rate</b>	-0.342870	0.069941	0.258984	1.000000

Table 4: Cholera case rate of Bangladesh from 1996 to 2000

<b>Country</b>	<b>Year</b>	<b>Number of reported cases of Cholera</b>	<b>Number of reported deaths from Cholera</b>	<b>Cholera case fatality rate</b>	<b>WHO Region</b>
Bangladesh	2000	1021	16	1.57	South-East Asia
Bangladesh	1999	3440	63	1.83	South-East Asia
Bangladesh	1998	1067	26	2.44	South-East Asia
Bangladesh	1997	1959	95	4.85	South-East Asia
Bangladesh	1996	418	0	0.00	South-East Asia

## CHAPTER 5

### CONCLUSIONS AND RECOMMENDATIONS

#### 5.1 Findings and Contributions

This research aims to extract some information and make a disease prediction through data analysis of Cholera case. We have accomplished this in our study using a variety of machine learning algorithms. We have used the most popular and powerful algorithms, such as Gradient Boosting. Gradient boosted machines (GBMs) are a widely used machine learning algorithm proven to be effective in various domains. Gradient Boosting works well with unbalanced data, such as real-time risk assessment. To the best of our knowledge, Cholera Disease Prediction has not previously used a Gradient boosting algorithm which we used first and achieved good accuracy at the same time.

We have observe that some papers worked on several algorithms such as, SARIMA, Markov chain Monte Carlo, SVMs, XGBoost, Cholera Artificial Learning Model (CALM), PCA, Genetic algorithms (GA), Clustering, dynamical Bayesian networks, Fuzzy Logic, object oriented design, Seasonal autoregressive integrated moving average (SARIMA) model, CART, Bayesian model averaging. In particularly for the case of SARIMA, Temporal clustering of Cholera was predicted. On the other hand, for the case of SVM, Cholera risks and better preparedness for public health management in the future was formulated. Also we have found that, XGBoost was considered in the previous study. According to our findings, there are many limitations in previous research. It is essential to test the hypothesis of a model because it is never possible to select a suitable model

without hypothesis testing of a model that has not been found in previous research. Previous research has not used biostatistics which is very important in the case of disease analysis that we are observing.

## **5.2 Recommendations for Future Works**

In the future, we will collect more datasets and do data analysis. Our findings will solve the problems and identify the Cholera case made by a robust machine learning model for model building. This research will also propose a pipeline for multiple disease prediction in future research. Although we have analysed some data by identifying Cholera disease, some more limitations are found in our research. We used some other algorithms besides Gradient boosting, but they did not get good accuracy because there was no variation in the dataset. The second problem I have faced is the availability of datasets. It is complicated to find a labelled dataset on this type of disease, and the sample size of the dataset we worked on was relatively small. We will extract confidential information also through some more statistical analysis like Biostatistics in Future Research.

In addition, cross connection on Seasonal Data and Demographic data to predict cholera prediction will be performed in the further study.

## REFERENCES

- Badkundri, R., Valbuena, V., Pinnamareddy, S., Cantrell, B., & Standeven, J. (2019). Forecasting the 2017-2018 Yemen cholera outbreak with machine learning. *arXiv preprint arXiv:1902.06739*.
- Daisy, S. S., Saiful Islam, A., Akanda, A. S., Faruque, A. S. G., Amin, N., & Jensen, P. K. M. (2020). Developing a forecasting model for cholera incidence in Dhaka megacity through time series climate data. *Journal of water and health*, 18(2), 207-223.
- Emch, M., Feldacker, C., Yunus, M., Streatfield, P. K., DinhThiem, V., & Ali, M. (2008). Local environmental predictors of cholera in Bangladesh and Vietnam. *The American journal of tropical medicine and hygiene*, 78(5), 823-832.
- González-Recio, O., Jiménez-Montero, J. A., & Alenda, R. (2013). The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. *Journal of Dairy Science*, 96(1), 614-624. doi:<https://doi.org/10.3168/jds.2012-5630>
- Ibrahim, N., Akhir, N. S. M., & Hassan, F. H. (2017). *Predictive analysis effectiveness in determining the epidemic disease infected area*. Paper presented at the AIP Conference Proceedings.
- kp, D. (2020). Cholera Dataset, No. of cases from different countries from 1949. Retrieved from <https://www.kaggle.com/imdevskp/cholera-dataset>
- Leo, J., Luhanga, E., & Michael, K. (2019). Machine Learning Model for Imbalanced Cholera Dataset in Tanzania. *The Scientific World Journal*, 2019, 9397578. doi:10.1155/2019/9397578
- Mubangizi, M., Mwebaze, E., & Quinn, J. A. (2009). Computational Prediction of Cholera Outbreaks. *Kampala. ICCIR*.
- Pasetto, D., Finger, F., Rinaldo, A., & Bertuzzo, E. (2017). Real-time projections of cholera outbreaks through data assimilation and rainfall forecasting. *Advances in Water Resources*, 108, 345-356.
- Singh, R., Singh, R., & Bhatia, A. (2018). Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics. *Int. J. Adv. Sci. Res*, 3(2), 19-24.
- WHO. (2021). Cholera  
Retrieved from <https://www.who.int/news-room/fact-sheets/detail/cholera>
- Chau, N. H. (2017, September). Enhancing Cholera Outbreaks Prediction Performance in Hanoi, Vietnam Using Solar Terms and Resampling Data. In *International Conference on Computational Collective Intelligence* (pp. 266-276). Springer, Cham.

Payment Ledger

N.B: To enjoy scholarship and tuition fee waiver undergraduate students must take at least 12 credits and postgraduate students must take at least 9 credits in each semester.

For more details please visit DIU website or <https://daffodilvarsity.edu.bd/scholarship/diu-scholarship>.

Student Info

Name: Roisujaman Shabab  
 ID: 161-35-1603  
 Email: roisujaman35-1603@diu.edu.bd

Payment Ledger Summary

Total Payable: **859800**  
 Total Paid: **859800**  
 Total Due: **0**  
 Total Other: **400**  
 Waiver / Scholarship calculation may vary the amounts

6/21/2021

Turnitin

Turnitin Originality Report

Processed on: 21-Jun-2021 15:24 +06  
 ID: 1609998892  
 Word Count: 5473  
 Submitted: 1

161-35-1603 By Roisujaman Shabab

Similarity Index  
**25%**

**Similarity by Source**  
 Internet Sources: 22%  
 Publications: 8%  
 Student Papers: 11%

4% match (Internet from 15-Mar-2020)  
[http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/3802/P15037%20%2822\\_%29.pdf?isAllowed=y&sequence=1](http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/3802/P15037%20%2822_%29.pdf?isAllowed=y&sequence=1)

3% match (student papers from 04-Apr-2018)  
 Class: Article 2018  
 Assignment: Journal Article  
 Paper ID: [940891790](#)

2% match (Internet from 17-Dec-2019)  
<http://www.allsciencejournal.com/archives/2018/vol3/issue2/3-2-15>

1% match (Internet from 11-Dec-2020)  
<https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html#:~:text=RMSE%20>

1% match ()  
[Daisy, Salima Sultana, Saiful Islam, A. K. M. et al. "Developing a forecasting model for cholera incidence in Dhaka megacity through time series climate data", 2020](#)

1% match (Internet from 13-Apr-2021)  
<https://www.mdpi.com/1660-4601/17/24/9378>

1% match (Internet from 12-Dec-2020)  
<https://data.humdata.org/dataset/who-data-for-ethiopia>

1% match (Internet from 27-Mar-2020)  
<https://www.science.gov/topicpages/t/travel+time+forecasting.html>

1% match (student papers from 14-Dec-2020)  
[Submitted to Westminster International University in Tashkent on 2020-12-14](#)

1% match (student papers from 21-Apr-2021)  
[Submitted to Monash University on 2021-04-21](#)

1% match (student papers from 24-Feb-2021)  
[Submitted to University of South Africa on 2021-02-24](#)

1% match (Internet from 21-Oct-2020)  
[https://www.researchgate.net/figure/Baseline-characteristics\\_tbl1\\_309225134](https://www.researchgate.net/figure/Baseline-characteristics_tbl1_309225134)

1% match (student papers from 17-May-2021)  
[Submitted to Gitam University on 2021-05-17](#)

1% match ()  
[Leo, Judith, Luhanga, Edith T., Michael, Kisangiri. "Machine Learning Model for Imbalanced Cholera Dataset in Tanzania.", 'Hindawi Limited', 2019](#)

1% match (Internet from 03-Oct-2020)  
<https://www.kaggle.com/imdevskp/cholera-dataset?select=data.csv>

1% match (publications)  
[Ashraful Islam Khan, Md Mahbubur Rashid, Md Taufiqul Islam, Mokibul Hassan Afrad et al. "Epidemiology of Cholera in Bangladesh: Findings From Nationwide Hospital-based Surveillance, 2014–2018", Clinical Infectious Diseases, 2020](#)

< 1% match (Internet from 08-Dec-2020)  
<https://www.mdpi.com/2076-3417/10/23/8634/htm>

< 1% match (Internet from 14-Oct-2016)  
<https://infoscience.epfl.ch/search?cc=Infoscience%2FResearch%2FENAC&of=hr&sf=year>

< 1% match (Internet from 11-Apr-2021)  
<http://ugspace.ug.edu.gh/bitstream/handle/123456789/26594/Development%20of%20Patient%20Record%20Management%20System%20for%20isAllowed=y&sequence=1>

< 1% match (student papers from 31-Jul-2019)  
[Submitted to uvt on 2019-07-31](#)

< 1% match (Internet from 20-Dec-2017)  
<http://dokumentix.ub.uni-siegen.de/opus/volltexte/2009/386/pdf/todt.pdf>

[https://www.turnitin.com/newreport\\_printview.asp?eq=1&eb=1&esm=10&oid=1609998892&sid=0&n=0&m=2&svr=54&r=42.098130051009775&lang=e...](https://www.turnitin.com/newreport_printview.asp?eq=1&eb=1&esm=10&oid=1609998892&sid=0&n=0&m=2&svr=54&r=42.098130051009775&lang=e...) 1/5

	<p>&lt; 1% match (Internet from 26-Jul-2020)  <a href="https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-020-01729-x">https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-020-01729-x</a></p>
	<p>&lt; 1% match (Internet from 11-Nov-2020)  <a href="https://reporting.unhcr.org/operation-reporting">https://reporting.unhcr.org/operation-reporting</a></p>
	<p>&lt; 1% match (student papers from 02-Nov-2015)  <a href="#">Submitted to University of Sydney on 2015-11-02</a></p>
	<p>&lt; 1% match (Internet from 17-Jun-2021)  <a href="https://iwaponline.com/jwh/article/18/2/207/72209/Developing-a-forecasting-model-for-cholera">https://iwaponline.com/jwh/article/18/2/207/72209/Developing-a-forecasting-model-for-cholera</a></p>
	<p>&lt; 1% match (student papers from 16-Dec-2020)  <a href="#">Submitted to Asia Pacific University College of Technology and Innovation (UCTI) on 2020-12-16</a></p>
	<p>&lt; 1% match (student papers from 07-Sep-2018) <a href="#">Submitted to The University of Manchester on 2018-09-07</a></p>
	<p>&lt; 1% match (Internet from 21-Mar-2021)  <a href="https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-0285-1">https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-0285-1</a></p>
	<p>&lt; 1% match (publications)  <a href="#">Najihah Ibrahim, Nur Shazwani Md. Akhir, Fadratul Hafinaz Hassan. "Predictive analysis effectiveness in determining the epidemic disease infected area", AIP Publishing, 2017</a></p>
	<p>&lt; 1% match (student papers from 31-May-2019)  <a href="#">Submitted to RDI Distance Learning on 2019-05-31</a></p>
	<p>&lt; 1% match (Internet from 20-Apr-2019)  <a href="http://events.science-japon.org/dlai17/doc/Osamu%20Sudoh%20-%20Social%20impact%20of%20AI.pdf">http://events.science-japon.org/dlai17/doc/Osamu%20Sudoh%20-%20Social%20impact%20of%20AI.pdf</a></p>
	<p>&lt; 1% match (Internet from 14-Dec-2020)  <a href="https://www.cdc.gov/cholera/diagnosis.html">https://www.cdc.gov/cholera/diagnosis.html</a></p>
	<p>&lt; 1% match (Internet from 10-Oct-2020)  <a href="https://www.springerprofessional.de/enhancing-cholera-outbreaks-prediction-performance-in-hanoi-viet/15066574">https://www.springerprofessional.de/enhancing-cholera-outbreaks-prediction-performance-in-hanoi-viet/15066574</a></p>
	<p>&lt; 1% match ()  <a href="#">Yaacob, Maslina. "Wide range analysis of ozone gas concentration in ultraviolet region", 2016</a></p>
<p>28</p>	<p>Analysis and Prediction of Cholera Disease using Machine Learning Algorithms By Roisujaman Shabab ID: 161-35-1603 <a href="#">This Report Presented in Fulfillment of the Requirements for the Degree of B.Sc. in Software Engineering Supervised By Ms. Farzana Sadia Assistant Professor DEPARTMENT OF SOFTWARE ENGINEERING DAFFODIL INTERNATIONAL UNIVERSITY Thesis Approval i Page Thesis Declaration ii ©Daffodil International University ACKNOWLEDGEMENT I am grateful to my creator for allowing me to complete this research work and learn so much. I am thankful to my research supervisor, Ms Farzana Sadia, to provide HASH(0x7f5ae9ca2b80) Mr. Md. Anwar Hossen, Assistant Professor, HASH(0x7f5ae9ca2f10). HASH(0x7f5ae9ca3378) who supported me throughout this venture. ii ©Daffodil International University HASH(0x7f5ae9c9f300)HASH(0x7f5ae9ca3030) on Analysis Effectiveness in Determining the Epidemic..... 4 2.3 Previous research on forecasting Cholera disease ..... 5 HASH(0x7f5ae9c9f588) Research Dataset ..... 12 iii ©Daffodil International University 3.2 Data Preprocessing ..... 12 3.3 Algorithm Selection HASH(0x7f5ae9c9e408).1.1 Experimental Result &amp; Evaluation Matrix..... 16 4.1.2 Exploratory Data Analysis 18 HASH(0x7f5ae9ca8bb8): Literature Review 8 Table 2: Accuray Report of Gradient Boosting Algorithm ..... 18 Table 3: Descriptive statistics of the dataset 24 Table 4: Correlation matrix of the Cholera disease attributes ..... 24 v ©Daffodil International University vi ©Daffodil International University vii ©Daffodil International University viii ©Daffodil International University LIST OF FIGURE S Figure 1: Architecture Diagram of Proposed Research ..... 12 Figure 2: Visualisation of data before cleaning ..... 13 Figure 3: Visualisation of data after cleaning 14 Figure 4: Exploratory data analysis for the top 10 countries 21 Figure 5: Finding the correlation between the independent variable..... 22 Figure 6: Investigating the average number of deaths from Cholera ..... 23 Figure 7: Investigating the Cholera status of Bangladesh 24 ix ©Daffodil International University LIST OF ABBREVIATIONS Abbreviation Explanation HASH(0x7f5ae9ca9dc8) ML Machine Learning DL Deep Learning x ©Daffodil International University ABSTRACT In this research, we have investigated Cholera disease and its fatality rate from previous years in terms of various countries. This disease is not new, yet researchers are currently utilising several approaches to detect the difficulties and extract hidden information from previous records. This study proposed an alternative solution in terms of Cholera disease. Several traditional machine learning algorithms are experimented with to analyse and predict the disorder from the existing dataset. The proposed research methodology has consisted of four phases, for instance, Research Dataset, Data Preprocessing and Algorithm Selection. Our investigation shows that the Gradient Boosting algorithm performs well in this type of dataset with an accuracy of 93%. We have discovered the R2 score, HASH(0x7f5ae9caa068)) are 0.932841%, 398.1827%, 158549.4724%, and 71.6621098%, respectively. We have carried out an exploratory data analysis where Cholera case data analysis of Bangladesh has been done from 1996 to 2000. In addition to the disease prediction, data analysis of different countries has been carried out, and correlations have been made through which the interrelationships of each indicator can be found very quickly. Keywords: Machine Learning (ML), RMSE, MSE, Gradient Boosting and Cholera Disease ix ©Daffodil International University CHAPTER 1 INTRODUCTION 1.1 Background Cholera disease is not only a global issue but also a historical issues in the world. HASH(0x7f5ae9ca9ff0). HASH(0x7f5ae9caa5f0) growth (©Daffodil International University) and responsible for the various problem of our social life. According to the recent statistics, the scientist reported that every year 1.3 to 4.0 million Cholera case fatality rate of every year is due to less awareness and lack of self-consciousness. In the context of Bangladesh, In 1991, a massive diarrhoea epidemic erupted in Bangladesh. To estimate the magnitude of Cholera</a></p>

during diarrhoea epidemics and to focus on Cholera-related public health problems in Bangladesh. Here, two types of people can be found in Bangladesh: the people who live in the urban area and another in the rural area. It is noticeable that those who lived in the urban area are aware than the people who stay in the village side. The risk of Cholera has

HASH(0x7f5ae9cad6a8). 1|Page 1.2 Motivation of the Research Cholera is a significant disease, and it is essential to do awareness and analysis among people. Medical data analysis is relatively difficult, so it is crucial to analyse it through computational techniques. Machine Learning (ML) can diagnose diseases more effectively at an earlier level, reducing the number of readmissions in hospitals and clinics. Technology has also advanced significantly in discovering and developing new medicines with tremendous potential for treating patients with complex conditions. So, it is a matter of great concern to bring an effective solution and analyse the difficulties. This proposed study's motivation is to utilise the ML approach predicting the Cholera case fatality rate from the existing dataset and Figure out meaningful insights. 1.3 Problem Statement We have found that the previous research has not been accomplished or adequately formulated. Most of the previous research has been done through statistical analysis, but very little work has been done on Cholera disease prediction using machine learning algorithms. We also noticed that no detailed pipeline on data processing had been provided in previous research, which has caught our attention. Many evaluation matrixes show how good the model is, which has not been shown in previous research. HASH(0x7f5ae9cad00) algorithm performs best in predicting Cholera disease? ? RQ2: Is there only one or multiple variables that positively influence the disease prediction? 2 @Daffodil International University

HASH(0x7f5ae9ca518) algorithm that performs well in terms of Cholera disease prediction on Cholera cases population demographic data. ? Analyzing Cholera Population on Demographic Data ? To extract meaningful insights from the previous Cholera cases records. ? To achieve promising accuracy by experimenting with various algorithms. ? To reduce the loss function as well as overfitting and assure the model's generalizability. 1.6 Research Scope We have carefully investigated the Cholera disease and predict using the previous dataset. Since we have researched the clinical aspect, it will be possible to identify the Cholera disease fatality rate through our model accurately as well as find out what kind of factor works with this disease. We have shown the correlation to the interrelation of each feature so that it will serve as a benchmark in the research community. 1.7 Thesis Organisation In the following, in second chapter we examined about different looks researches done on the same subject including the research gap. In chapter three, we examined HASH(0x7f5ae9cae1d0) discussed the findings, limitations and HASH(0x7f5ae9caf140) research (Emch et al., 2008), the environmental conditions have been linked to Cholera and thus may help predict the disease. Compared to satellite-derived and in-situ ecological time-series data in Bangladesh and Vietnam, temporal Cholera distributions are compared to satellite-derived and in-situ ecological time-series data to investigate the links between Cholera and the local climate. In Bangladesh, ordered probit models are used to look at associations, while in Vietnam, probit models are used to look at associations at two different locations. Increased levels of ocean chlorophyll are linked to a rise in the severity of Cholera in Bangladesh. Increases in sea surface temperature. In Hue, Vietnam, the weather has the most significant impact, while in Nha Trang, Vietnam, river height has a significant effect. 2.2 Previous research on Analysis Effectiveness in Determining the Epidemic Epidemic disease outbreaks have prompted today's population to become increasingly concerned with infectious disease control, prevention, and management strategies to reduce disease spread and infected areas. For the contagious disease countermeasure and prediction study, the backpropagation approach was used in this research (Ibrahim, Akhir, & Hassan, 2017). Machine learning can evaluate the HASH(0x7f5ae9caf5c0) approach, which encourages artificial intelligence HASH(0x7f5ae9caf368). The feature collection of HASH(0x7f5ae9caf8a8) classification technique. 4 @Daffodil International University

HASH(0x7f5ae9caf350) study (Pasetto, Finger, Rinaldo, & Bertuzzo, 2017) introduces two major Cholera modelling innovations: using HASH(0x7f5ae9cb2d80). A benchmark was developed by HASH(0x7f5ae9cb3038). Their findings indicate that the assimilation protocol with sequential parameter updates outperforms Markov chain Monte Carlo calibration schemes. In (Singh, Singh, & Bhatia, 2018), the authors addressed that datasets available on the internet containing helpful information are known as sentiment analysis (SA). In the field of computer science, machine learning techniques play a significant role. The study of patterns and the creation of HASH(0x7f5ae9cb3590) is HASH(0x7f5ae9cb33b0). There have been studies in the recent past. HASH(0x7f5ae9cb3608). The importance HASH(0x7f5ae9cb4680) using HASH(0x7f5ae9cb48d8). HASH(0x7f5ae9cb4ea8) and 2013 due to the auto-regressive nature of Cholera and its seasonal 5 @Daffodil International University

behavioural trends (Daisy et al., 2020). Since HASH(0x7f5ae9cb7cf0). Relative humidity (r 14 HASH(0x7f5ae9cb8398) 14 -0.23) were all found to have a weak relationship. A 7% rise in Cholera incidence (p 0.001) was observed HASH(0x7f5ae9cb8548). MVM, on the other hand, outperformed SVM (AIC 14 15, BIC 14 36) in terms of efficiency. 2.3 Previous research on forecasting Cholera disease The authors (Daisy et al., 2020) present HASH(0x7f5ae9cb8818), which has a population of more than 200 million people, using critical climate variables derived from atmospheric, terrestrial, and oceanic satellites. HASH(0x7f5ae9cb8aa0) coast of India from 2010 to 2018. With HASH(0x7f5ae9cb9788), the random forest classifier model correctly identifies 89.5 percent of outbreaks. Seasons and coastal locations revealed spatial- temporal trends in the model's results. The authors (Badkundri, Valbuena, Pinnamareddy, Cantrell, & Standeven, 2019) proposed HASH(0x7f5ae9cb9998) period ranging from two weeks to two months. CALM uses HASH(0x7f5ae9cb9db8) experiences from different time frames to create a novel machine learning approach. CALM may also teach complex nonlinear relations that appear in epidemiological phenomena thanks to 6 @Daffodil International University

machine learning and comprehensive function engineering. In a real-world simulation, CALM can predict Cholera incidence 2 weeks to two months ahead of time with a margin of 5 Cholera cases per 10,000 people. The study (Leo, Luhanga, & Michael, 2019) proposed HASH(0x7f5ae9cb9bf0) paper. The dataset's sampling equilibrium and dimensionality were restored using the HASH(0x7f5ae9cb3b8) performance of the seven models was also assessed using HASH(0x7f5ae9cb9980). XGBoost classifier was chosen as the best model for the analysis based on the Wilcoxon sign-rank test results and model features. Overall, the findings helped us better understand the critical functions of machine learning techniques in healthcare data. Early identification of Cholera cases is critical for limiting the severity and length of an outbreak. However, by the time Cholera cases are reported, an epidemic could already be well underway. This paper explains how to use data from various sources and machine learning techniques to forecast the likelihood of Cholera outbreaks in different areas over time. The authors examine regions with similar Cholera prevalence dynamics over time using data collected across 80 Ugandan regions. They then develop a probabilistic model to forecast possible Cholera cases and explain how it works (Mubangizi, Mwebaze, & Quinn, 2009). The authors (Chau et al., 2017) of this paper suggested a new approach to improve cholera outbreak prediction efficiency in Hanoi, Vietnam. Solar terminology, training data resampling, and classification methods are all included in the new approach. The results of the experiments HASH(0x7f5ae9cb2c0) results in (AUC) with balanced sensitivity and specificity. 7 @Daffodil International University

2.4 Research Gap Based on the review of the above research, it can be said that most of the research has been completed with analysis but using machine learning technology in the context of Disease Prediction and Bangladesh. By looking at the above literature review, it is obvious that majority of the research concentrated on statistical analysis or time series data handling for forecasting the outbreak, but to the best of our knowledge, there is not adequate study carried out in terms of cholera disease cases fatality rate prediction using machine learning approach. In this research, we have come up with an effective solution for combating the issues. Therefore, by using our model it is possible to predict the fatality rate of cholera disease through the computational approach. On the most the cases, researcher often use their own strategy for solving this issues, but it is a matter of great concern that analyzing the disease in the context of Bangladesh is crucial that cannot be found beforehand in the previous study. 2.5 Summary A variety of machine learning algorithms have been used in our research, and Cholera Disease has been identified by choosing the best algorithm from there. We have extracted information from the previous data through data analysis technique in Bangladesh, which has not been extensively completed during the existing research. 8 @Daffodil International University

CHAPTER 3 RESEARCH METHODOLOGY Figure 1 illustrates the architecture diagram of the proposed research. By looking at Figure 1, it can be clearly observed that the architecture is consisted of several phases, for instance, Data collection, Data Preprocessing, Data Visualization, Extract Information, Algorithm Selection, Model Evaluating and Classifier. Hence, in this

way, the disease has been classified and analyse the hidden information; however, the entire research is concentrated by following this procedure. Figure 1: Architecture Diagram of Proposed Research 3.1 Research Dataset We have used this dataset from the online repository (kp, 2020). This dataset HASH(0x7f5ae9cbd3f8). The number of reported deaths from Cholera can be found around 2492, and the Cholera case fatality rate is around 2492. [9 @Daffodil International University](#) 3.2 Data Preprocessing The main HASH(0x7f5ae9cbd920). In this step, we have removed the null value from our dataset since it can be responsible for many issues; however, it has to be solved in order to feed the data into the machine learning algorithm. Figure 2 and Figure 3 shows the noise linked dataset illustration and cleaning dataset visualisation. Figure 2: Visualisation of the noisy data before cleaning [10 @Daffodil International University](#) Figure 3: Visualisation of the data after cleaning 3.3 Algorithm Selection We have used seven popular HASH(0x7f5ae9cbdd10)), Adaptive Boosting, Gradient Boosting, K nearest neighbors (KNN) and Decision Tree. In this section, the Gradient Boosting algorithm will have been interpreted because we have found satisfactory accuracy on this approach in terms of the Cholera disease prediction. Gradient Boosting: The Gradient boosting algorithm also follows the sequential ensemble learning method. Through loss optimisation in this way, WikiLearners gradually became better than its previous weak learners. For example, the 2nd weak learner is better than the 1st. The 3rd weak learner is better than the 2nd, so as the weak learner periodicity increases, the amount of [11 @Daffodil International University](#) error in the model decreases, and the model becomes a stronger learner. The Gradient boosting algorithm works relatively well for regression type problems (González-Recio, Jiménez- Montero, & Alenda, 2013). The difference between Gradient boosting and Ada boosting is that in Adaptive boosting, error is gradually reduced by updating the weight of the wrong predictive air sample. In Gradient, boosting the loss function is optimised, and each loss is optimised. The amount of error also decreases. To optimise this loss function, each weak learner changes its alternative weak learner model so that the next weak learner is better than the previous one. On the other hand, Gradient boosting consists of 3 components, weak learner, loss function optimisation and additive model. The following equation (1),

(2), (3), (4), (5) and (6) are shown mathematically the working procedure of the Gradient boosting algorithm. 1: Initialise the function estimate  $f_0 = 0$ ,  $\sigma = \infty$ ,  $R = \{ \arg \min_{\tau} \sum_{i=1}^n L(\tau(x_i)) \}$  (1) -1 Step2: For each iteration  $m = 1, \dots, M$ : Calculate pseudo-residuals,  $r_m = -\frac{\partial L(f)}{\partial f}$ , for  $f = f_{m-1}$ , Add a new function  $\tau_m$  (R can be any model, but here we are using decision trees) as regression on pseudo-residuals  $r_m$ ,  $f_m = f_{m-1} + \tau_m$  (2) (3) iii. Find optimal coefficient at  $\tau_m$  regarding initial loss function  $L = \arg \min_{\alpha} \sum_{i=1}^n L(\tau_m(x_i) + f_{m-1}(x_i))$  (4)  $\alpha = \frac{1}{\sum_{i=1}^n \exp(-\alpha \tau_m(x_i))}$  Update current approximation  $f_m = f_{m-1} + \alpha \tau_m$  (5)  $\sigma = \sum_{m=1}^M \tau_m(x_i)$  (6) Step 3: The final GBM model will be the sum of the initial constant and all the subsequent function updates  $f = \sum_{m=0}^M \tau_m$  (7) (8) [13 @Daffodil International University](#) CHAPTER 4 RESULTS AND DISCUSSIONS 4.1 Data Analysis & Experimental Result The result analysis section is subdivided into three segments: Experimental Result & Model Evaluation, Exploratory Data Analysis and Comparative Analysis. 4.1.1 Experimental Result To explain how well the model is doing in its predictions, HASH(0x7f5ae9cbea58) developing machine learning models. The evaluation metrics differ depending on the type of problem. HASH(0x7f5ae9cbee60)'s goals only include HASH(0x7f5ae9cbf0e8) based on criteria. We experimented with the linear regression model in this dataset, but we did not find satisfactory accuracy; in the same dataset, the Gradient boosting algorithm has been applied, and good accuracy has been achieved. Table HASH(0x7f5ae9cbf868)HASH(0x7f5ae9cc2530) efficiency. Figure 4 shows the accuracy graphs in terms of the algorithms that have been experimented while conducting this research towards Cholera diseases prediction. [14 @Daffodil International University](#) Table 1: Accuracy report of the Gradient boosting algorithm MAE % MSE% RMSE% Accuracy% R2% 71.6621098 158549.4724 398.1827 93.284 0.932841 ACCURACY 93 53 40 38 36 41 41 18 5 Figure 4: Accuracy Graphs of the algorithms that have been applied 4.1.2 Evaluation Matrix This section has been divided into several phases: HASH(0x7f5ae9cc2980). A. HASH(0x7f5ae9cc27a0) square root of MSE divided by the error rate. HASH(0x7f5ae9cc4160). D. R-squared (Coefficient of determination): It represents the coefficient indicating HASH(0x7f5ae9cc3f20). The equation (7), (8), (9) and (10) are used to find out the prediction error and model efficiency.  $\sum_{i=1}^n |y_i - \hat{y}_i| = 1$  (7)  $\sum_{i=1}^n (y_i - \bar{y})^2 = 1$  (8)  $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = 1$  (9)  $\sum_{i=1}^n (y_i - \bar{y})^2 = 1$  (10) Where  $\hat{y}_i$ -predicted value of  $y_i$ -mean value of [16 @Daffodil International University](#)

4.1.3 Exploratory Data Analysis This section analyses the Cholera case on top of different countries. Without analysing HASH(0x7f5ae9cc4628)HASH(0x7f5ae9cc4a60) may help to detect Cholera. While the treatment HASH(0x7f5ae9cc7740) the risk of a widespread outbreak. Bangladesh only has data from the year 2010 onwards. Despite a slight rise in the number of Cholera cases, we can see that Bangladesh has made progress in its Cholera war. This data set has some problems with data quality, such as missing values, data form mismatches, and invalid number input. Which ones were already corrected during the analysis? We can see that the number of outliers is minimal. We also discovered that not all countries have data for all years. The most significant number of confirmed deaths in 2016 was in the HASH(0x7f5ae9cc7cb0), Yemen, South Sudan, Kenya, Malawi, Nigeria, and the Dominican Republic. The countries with the most Cholera cases in 2016 were HASH(0x7f5ae9cc7e48), Kenya, South Sudan, Malawi, the Dominican Republic, and Mozambique. The countries with the highest fatality rates in 2016 were Niger, Congo, Zimbabwe, Nigeria, Angola, Somalia, the HASH(0x7f5ae9cc7728) total number of outbreaks, average death, and fatality rate have all decreased over time, but Cholera disease plagues a few countries. To eliminate Cholera, we need to focus more on these countries. The details sequence and consequence are shown separately in Figure 5,6,7, and 8. [17 @Daffodil International University](#) Figure 5: Exploratory data analysis for the top 10 countries with the most amount of Cholera disease. [18 @Daffodil International University](#) Figure 6: Finding the correlation between the independent variables in terms of Cholera disease. [19 @Daffodil International University](#) Figure 7: Investigating the average number of deaths from Cholera from 1950 to 2010 [20 @Daffodil International University](#) Figure 8: Investigating the Cholera status of Bangladesh from 1973 to 2000 The total number of outbreaks, average death, and fatality rate have all decreased over time, but Cholera disease plagues a few countries. To eliminate Cholera, we need to focus more on these countries. The details sequence and consequence are shown separately in Figure 2,3,4, and 5. Table 2: Statistical analysis & measurements of the dataset Year HASH(0x7f5ae9cc8520) Count 2492.000000 2492.000000 2492.000000 2492.000000 Mean 1992.343098 3684.060193 360.033708 5.459960 Std 14.834151 14840.198322 3484.892806 15.211705 Min 1949.000000 0.000000 0.000000 0.000000 25% 1981.000000 8.750000 0.000000 50% 1994.000000 228.500000 5.000000 1.300000 75% 2004.000000 1847.750000 53.250000 4.912500 max 2016.000000 340311.000000 124227.000000 450.000000 [21 @Daffodil International University](#) Table 3: Correlation matrix of the Cholera disease attributes Year HASH(0x7f5ae9ccd940) Year 1.000000 -0.011789 -0.184387 -0.342870 HASH(0x7f5ae9cc9400) -0.011789 1.000000 0.442026 0.069941 HASH(0x7f5ae9cc9340) -0.184387 0.442026 1.000000 0.258984 Cholera case fatality rate -0.342870 0.069941 0.258984 1.000000 Table 4: Cholera case rate of Bangladesh from 1996 to 2000 Country Year HASH(0x7f5ae9cc4208) WHO Region Bangladesh 2000 1021 16 1.57 South-East Asia Bangladesh 1999 3440 63 1.83 South-East Asia Bangladesh 1998 1067 26 2.44 South-East Asia Bangladesh 1997 1959 95 4.85 South-East Asia Bangladesh 1996 418 0 0.00 South-East Asia [22 @Daffodil International University](#) HASH(0x7f5ae9cd0638) aims to extract some information and make a disease prediction through data analysis of Cholera case. We have accomplished this in our study using a variety of machine learning algorithms. We have used the most popular and powerful algorithms, such as Gradient Boosting. HASH(0x7f5ae9cd0d40) in various domains. Gradient Boosting works well with unbalanced data, such as real-time risk assessment. To the best of our knowledge, Cholera Disease Prediction has not previously used a Gradient boosting algorithm which we used first and achieved good accuracy at the same time. We have observe that some papers worked on several algorithms such as, SARIMA, Markov chain Monte Carlo, SVMs, XGBoost, Cholera Artificial Learning Model (CALM), PCA, Genetic algorithms (GA), Clustering, dynamical Bayesian networks, Fuzzy Logic, object oriented design, Seasonal autoregressive integrated moving average (SARIMA) model, CART, Bayesian model averaging. In particularly for the case of SARIMA, Temporal clustering of Cholera was predicted. On the other hand, for the case of SVM, HASH(0x7f5ae9cd10b8) was formulated. Also we have found that, XGBoost was considered in a previous study. According to our findings, there are many limitations in previous research. It is essential to test the hypothesis of a model because it is never possible to select a suitable model [23 @Daffodil International University](#) without hypothesis testing

of a model that has not been found in previous research. Previous research has not used biostatistics which is very important in the case of disease analysis that we are observing. 5.2 Recommendations for Future Works In the future, we will collect more datasets and do data analysis. Our findings will solve the problems and identify the Cholera case made by a robust machine learning model for model building. This research will also propose a pipeline for multiple disease prediction in future research. Although we have analysed some data by identifying Cholera disease, some more limitations are found in our research. We used some other algorithms besides Gradient boosting, but they did not get good accuracy because there was no variation in the dataset. The second problem I have faced is the availability of datasets. It is complicated to find a labelled dataset on this type of disease, and the sample size of the dataset we worked on was relatively small. We will extract confidential information also through some more statistical analysis like Biostatistics in Future Research. In addition, cross connection on Seasonal Data and Demographic data to predict cholera prediction will be performed in the further study. 24 ©Daffodil International University

REFERENCES

Badkundri, R., Valbuena, V., Pinnamareddy, S., Cantrell, B., & Standeven, J. (2019). Forecasting the 2017-2018 Yemen cholera outbreak with machine learning. arXiv preprint arXiv:1902.06739.

Daisy, S. S., Saiful Islam, A., Akanda, A. S., Faruque, A. S. G., Amin, N., & Jensen, P. K. M. (2020). Developing a forecasting model for cholera incidence in Dhaka megacity through time series climate data. *Journal of water and health*, 18(2), 207-223.

Emch, M., Feldacker, C., Yunus, M., Streatfield, P. K., DinhThiem, V., & Ali, M. (2008). Local environmental predictors of cholera in Bangladesh and Vietnam. *The American journal of tropical medicine and hygiene*, 78(5), 823-832.

González-Recio, O., Jiménez-Montero, J. A., & Alenda, R. (2013). The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. *Journal of Dairy Science*, 96(1), 614-624. doi:<https://doi.org/10.3168/jds.2012-5630>

Ibrahim, N., Akhir, N. S. M., & Hassan, F. H. (2017). Predictive analysis effectiveness in determining the epidemic disease infected area. Paper presented at the AIP Conference Proceedings.

kp, D. (2020). Cholera Dataset, No. of cases from different countries from 1949. Retrieved from <https://www.kaggle.com/imdevskp/cholera-dataset>

Leo, J., Luhanga, E., & Michael, K. (2019). Machine Learning Model for Imbalanced Cholera Dataset in Tanzania. *The Scientific World Journal*, 2019, 9397578. doi:10.1155/2019/9397578

Mubangizi, M., Mwebaze, E., & Quinn, J. A. (2009). Computational Prediction of Cholera Outbreaks. Kampala. ICCIR.

Pasetto, D., Finger, F., Rinaldo, A., & Bertuzzo, E. (2017). Real-time projections of cholera outbreaks through data assimilation and rainfall forecasting. *Advances in Water Resources*, 108, 345-356.

Singh, R., Singh, R., & Bhatia, A. (2018). Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics. *Int. J. Adv. Sci. Res*, 3(2), 19-24.

WHO. (2021). Cholera Retrieved from <https://www.who.int/news-room/fact-sheets/detail/cholera>

Chau, N. H. (2017, September). Enhancing Cholera Outbreaks Prediction Performance in Hanoi, Vietnam Using Solar Terms and Resampling Data. In *International Conference on Computational Collective Intelligence* (pp. 266-276). Springer, Cham. 25 ©Daffodil International University 26 ©Daffodil International University

[https://www.turnitin.com/newreport\\_printview.asp?eq=1&eb=1&esm=10&oid=1609998892&sid=0&n=0&m=2&svr=54&r=42.098130051009775&lang=e...](https://www.turnitin.com/newreport_printview.asp?eq=1&eb=1&esm=10&oid=1609998892&sid=0&n=0&m=2&svr=54&r=42.098130051009775&lang=e...)

5/5